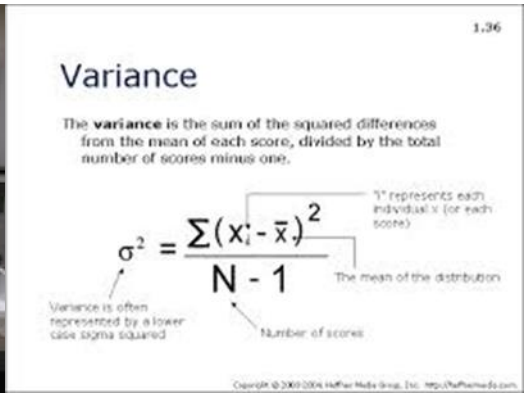
EXPERIMENTAL DESIGN AND THE VARIANCE

SESSION 9



Variance and Standard Deviation

Deviations about the mean of a population is the basis for most of the statistical tests we will learn. Since we are measuring how widely a set of scores is dispersed about the mean we are measuring variability. We can calculate the deviations about the mean, and express it as variance or standard deviation. It is very important to have a firm grasp of this concept because it will be a central concept throughout the course.

Both variance and standard deviation measures variability within a distribution. Standard deviation is a number that indicates how much, on average, each of the values in the distribution deviates from the mean (or center) of the distribution. Keep in mind that variance measures the same thing as standard deviation (dispersion of scores in a distribution). Variance, however, is the average squared

deviations about the mean. Thus, variance is the square of the standard deviation.

In terms of quality of goods/services, It is important to know that higher variation means lower quality. Measuring the size of variation and its source is the statistician's job, while fixing it is the job of the engineer or the manager. Quality products and services have low variation.

True experimental design is regarded as the most accurate form of experimental research, in that it tries to

For some of the physical sciences, such as physics, chemistry and geology, they are standard and commonly used. For social sciences, psychology and biology, they can be a little more difficult to set up.

For an experiment to be classed as a true experimental design, it must fit all of the following criteria.

- The sample groups must be assigned randomly.
- There must be a viable control group.
- Only one variable can be manipulated and tested. It is possible to test more than one, but such experiments and their statistical analysis tend to be cumbersome and difficult.
- The tested subjects must be randomly assigned to either control or experimental groups.

Advantages

The results of a true experimental design can be statistically analyzed and so there can be little argument about the results.

It is also much easier for other researchers to replicate the experiment and validate the results.

For physical sciences working with mainly numerical data, it is much easier to manipulate one variable, so true experimental design usually gives a yes or no answer.

Disadvantages

Whilst perfect in principle, there are a number of problems with this type of design. Firstly, they can be almost too perfect, with the conditions being under complete control and not being representative of real world conditions.

For psychologists and behavioral biologists, for example, there can never be any guarantee that a human or living organism will exhibit 'normal' behavior under experimental conditions.

True experiments can be too accurate and it is very difficult to obtain a complete rejection or acceptance of a hypothesis because the standards of proof required are so difficult to reach.

True experiments are also difficult and expensive to set up. They can also be very impractical.

While for some fields, like physics, there are not as many variables so the design is easy, for social sciences and biological sciences, where variations are not so clearly

defined it is much more difficult to exclude other factors that may be affecting the manipulated variable.

Summary

True experimental design is an integral part of science, usually acting as a final test of a hypothesis. Whilst they can be cumbersome and expensive to set up, literature reviews, qualitative research and descriptive research can serve as a good precursor to generate a testable hypothesis, saving time and money.

Whilst they can be a little artificial and restrictive, they are the only type of research that is accepted by all disciplines as statistically provable

Experimental Design

Experimental designs are often touted as the most "rigorous" of all research designs or, as the "gold standard" against which all other designs are judged. In one sense, they probably are. If you can implement an experimental design well (and that is a big "if" indeed), then the experiment is probably the strongest design with respect to internal validity. Why? Recall that internal validity is at the center of all causal or cause-effect inferences. When you want to determine whether some program or treatment causes some outcome or outcomes to occur, then you are

interested in having strong internal validity. Essentially, you want to assess the proposition:

If X, then Y

or, in more colloquial terms:

If the program is given, then the outcome occurs

Unfortunately, it's not enough just to show that when the program or treatment occurs the expected outcome also happens. That's because there may be lots of reasons, other than the program, for why you observed the outcome. To really show that there is a causal relationship, you have to simultaneously address the two propositions:

If X, then Y

and

If not X, then not Y

Or, once again more colloquially:

If the program is given, then the outcome occurs

and

If the program is not given, then the outcome does not occur

If you are able to provide evidence for both of these propositions, then you've in effect isolated the program from all of the other potential causes of the outcome. You've shown that when the program is present the outcome occurs and when it's not present, the outcome doesn't occur. That points to the causal effectiveness of the program.

Think of all this like a fork in the road. Down one path, you implement the program and observe the outcome. Down the other path, you don't implement the program and the outcome doesn't occur. But, how do we take both paths in the road in the same study? How can we be in two places at once? Ideally, what we want is to have the same conditions -- the same people, context, time, and so on --

and see whether when the program is given we get the outcome and when the program is not given we don't. Obviously, we can never achieve this hypothetical situation. If we give the program to a group of people, we can't simultaneously not give it! So, how do we get out of this apparent dilemma?

Perhaps we just need to think about the problem a little differently. What if we could create two groups or contexts that are as similar as we can possibly make them? If we could be confident that the two situations are comparable, then we could administer our program in one (and see if the outcome occurs) and not give the program in the other (and see if the outcome doesn't occur). And, if the two contexts are comparable, then this is like taking both forks in the road simultaneously! We can have our cake and eat it too, so to speak.

That's exactly what an experimental design tries to achieve. In the simplest type of experiment, we create two groups that are "equivalent" to each other. One group (the program or treatment group) gets the program and the other group (the comparison or control group) does not. In all other respects, the groups are treated the same. They have similar people, live in similar contexts, have similar backgrounds, and so on. Now, if we observe differences in outcomes between these two groups, then the differences

must be due to the only thing that differs between them -- that one got the program and the other didn't.

OK, so how do we create two groups that are "equivalent"? The approach used in experimental design is to assign people randomly from a common pool of people into the two groups. The experiment relies on this idea of random assignment to groups as the basis for obtaining two groups that are similar. Then, we give one the program or treatment and we don't give it to the other. We observe the same outcomes in both groups.

The key to the success of the experiment is in the random assignment. In fact, even with random assignment we never expect that the groups we create will be exactly the same. How could they be, when they are made up of different people? We rely on the idea of probability and assume that the two groups are "probabilistically equivalent" or equivalent within known probabilistic ranges.

So, if we randomly assign people to two groups, and we have enough people in our study to achieve the desired probabilistic equivalence, then we may consider the experiment to be strong in internal validity and we probably have a good shot at assessing whether the program causes the outcome(s).

But there are lots of things that can go wrong. We may not have a large enough sample. Or, we may have people who refuse to participate in our study or who drop out part way through. Or, we may be challenged successfully on ethical grounds (after all, in order to use this approach we have to deny the program to some people who might be equally deserving of it as others). Or, we may get resistance from the staff in our study who would like some of their "favorite" people to get the program. Or, they mayor might insist that her daughter be put into the new program in an educational study because it may mean she'll get better grades.

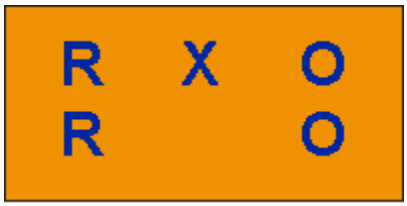
The bottom line here is that experimental design is intrusive and difficult to carry out in most real world contexts. And, because an experiment is often an intrusion, you are to some extent setting up an artificial situation so that you can assess your causal relationship with high internal validity. If so, then you are limiting the degree to which you can generalize your results to real contexts where you haven't set up an experiment. That is, you have reduced your external validity in order to achieve greater internal validity.

In the end, there is just no simple answer (no matter what anyone tells you!). If the situation is right, an experiment can

be a very strong design to use. But it isn't automatically so. My own personal guess is that randomized experiments are probably appropriate in no more than 10% of the social research studies that attempt to assess causal relationships.

Experimental design is a fairly complex subject in its own right. I've been discussing the simplest of experimental designs -- a two-group program versus comparison group design. But there are lots of experimental design variations that attempt to accomplish different things or solve different problems. In this section you'll explore the basic design and then learn some of the principles behind the major variations.

Two-Group Experimental Designs



The simplest of all experimental designs is the two-group posttest-only randomized experiment. In design notation, it has two lines -- one for each group -- with an R at the beginning of each line to indicate that the groups were randomly assigned. One group gets the treatment or program (the X) and the other group is the comparison group and doesn't get the program (note that this you could alternatively have the comparison group receive the standard or typical treatment, in which case this study would be a relative comparison).

history	✓
maturation	✓
testing	✓
instrumentation	✓
mortality	✓
regression to the mean	✓
selection	✓
selection-history	✓
selection- maturation	✓
selection- testing	✓
selection- instrumentation	✓
selection- mortality	✗
selection- regression	✓
diffusion or imitation	✗
compensatory equalization	✗
compensatory rivalry	✗
resentful demoralization	✗

Notice that a pretest is not required for this design. Usually we include a pretest in order to determine whether groups are comparable prior to the program, but because we are using random assignment we can assume that the two groups are probabilistically equivalent to begin with and the pretest is not required (although you'll see with covariance designs that a pretest may still be desirable in this context).

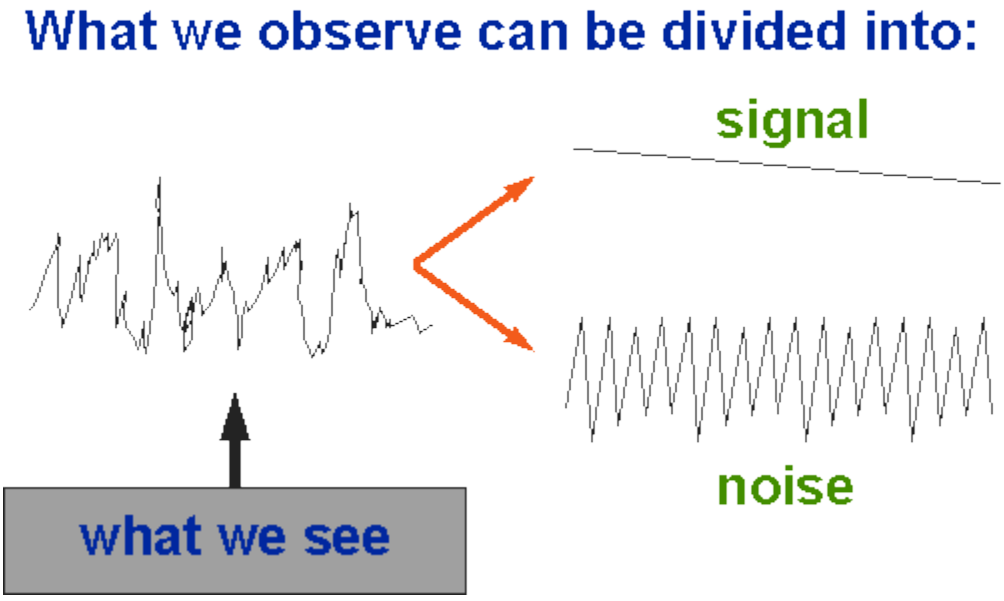
In this design, we are most interested in determining whether the two groups are different after the program. Typically we measure the groups on one or more measures (the Os in notation) and we compare them by testing for the differences between the means using a t-test or one way Analysis of Variance (ANOVA).

The posttest-only randomized experiment is strong against the single-group threats to internal validity because it's not a single group design! (Tricky, huh?) It's strong against the all of the multiple-group threats except for selection-mortality. For instance, it's strong against selection-testing and selection-instrumentation because it doesn't use repeated measurement. The selection-mortality threat is especially salient if there are differential rates of dropouts in the two groups. This could result if the treatment or program is a noxious or negative one (e.g., a painful medical procedure

like chemotherapy) or if the control group condition is painful or intolerable. This design is susceptible to all of the social interaction threats to internal validity. Because the design requires random assignment, in some institutional settings (e.g., schools) it is more likely to utilize persons who would be aware of each other and of the conditions they've been assigned to.

The posttest-only randomized experimental design is, despite its simple structure, one of the best research designs for assessing cause-effect relationships. It is easy to execute and, because it uses only a posttest, is relatively inexpensive. But there are many variations on this simple experimental design. You can begin to explore these by looking at how we classify the various experimental designs

Classifying Experimental Designs



Although there are a great variety of experimental design variations, we can classify and organize them using a simple signal-to-noise ratio metaphor. In this metaphor, we assume that what we observe or see can be divided into two components, the

signal and the noise (by the way, this is directly analogous to the true score theory of measurement). The figure, for instance, shows a time series with a slightly downward slope. But because there is so much variability or noise in the series, it is difficult even to detect the downward slope. When we divide the series into its two components, we can clearly see the slope.

In most research, the signal is related to the key variable of interest -- the construct you're trying to measure, the program or treatment that's being implemented. The noise consists of all of the random factors in the situation that make it harder to see the signal -- the lighting in the room, local distractions, how people felt that day, etc. We can construct a ratio of these two by dividing the signal by the noise. In research, we want the signal to be high relative to the noise. For instance, if you have a very powerful treatment or program (i.e., strong signal) and very good measurement (i.e., low noise) you will have a better chance of seeing the effect of the program than if you have either a strong program and weak measurement or a weak program and strong measurement.

$$\frac{\text{signal}}{\text{noise}}$$

With this in mind, we can now classify the experimental designs into two categories: **signal enhancers** or **noise reducers**. Notice that doing either of these things -- enhancing signal or reducing noise -- improves the quality of the research. The *signal-enhancing experimental designs* are called the factorial designs. In these designs, the focus is almost entirely on the setup of the program or treatment, its components and its major dimensions. In a typical factorial

design we would examine a number of different variations of a treatment.

There are two major types of *noise-reducing experimental designs*: covariance designs and blocking designs. In these designs we typically use information about the makeup of the sample or about pre-program variables to remove some of the noise in our study

In probability theory and statistics, variance measures how far a set of numbers is spread out. A variance of zero indicates that all the values are identical. Variance is always non-negative: a small variance indicates that the data tends to be very close to the mean (expected value) and hence to each other, while a high variance indicates that the data is very spread out around the mean and from each other.

An equivalent measure is the square root of the variance, called the standard deviation. The standard deviation has the same dimension as the data, and hence is comparable with deviations of the mean.

The variance is one of several descriptors of a probability distribution. In particular, the variance is one of the moments of a distribution. In that context, it forms part of a systematic approach to distinguishing between probability distributions. While other such approaches have been developed, those

based on moments are advantageous in terms of mathematical and computational simplicity.

The variance is a parameter that describes, in part, either the actual probability distribution of an observed population of numbers, or the theoretical probability distribution of a sample (a not-fully-observed population) of numbers. In the latter case, a sample of data from such a distribution can be used to construct an estimate of its variance: in the simplest cases this estimate can be the sample variance.

Statistical variance gives a measure of how the data distributes itself about the mean or expected value. Unlike range that only looks at the extremes, the variance looks at all the data points and then determines their distribution.